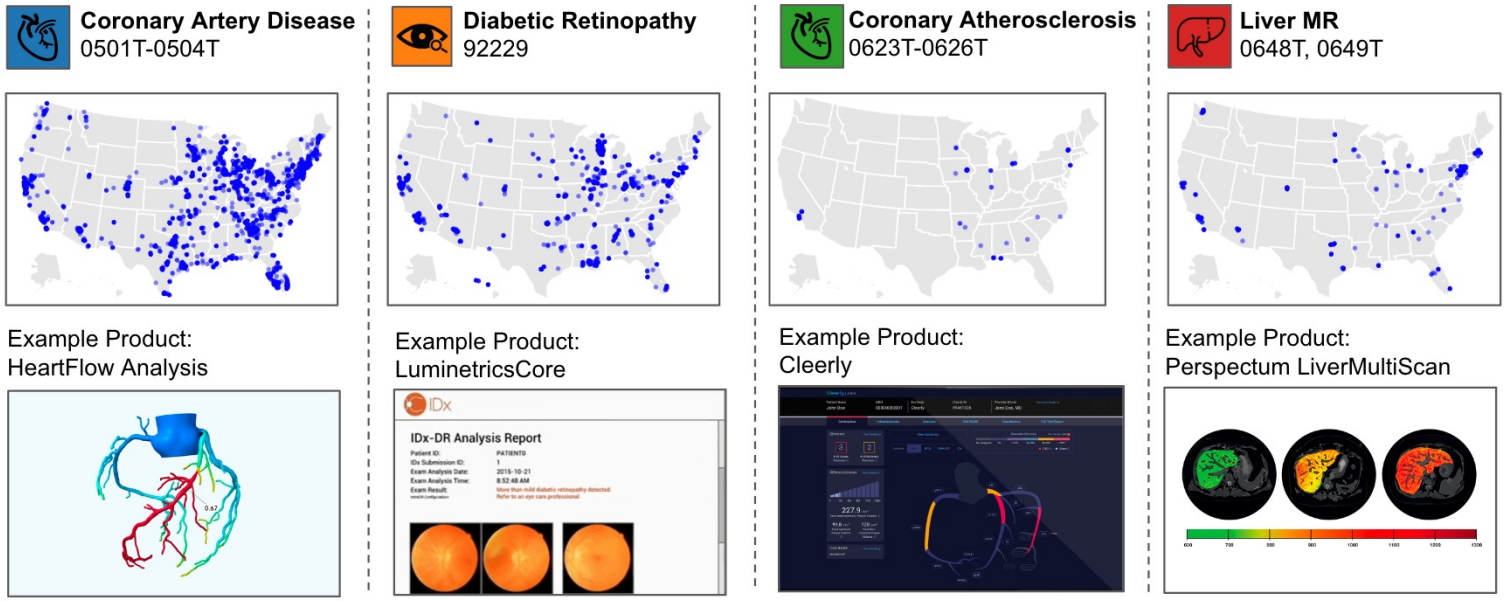
Stanford University

CHAN ZUCKERBERG BIOHUB

How are medical AI devices evaluated, updated and deployed?

James Zou



www.james-zou.com

jamesz@stanford.edu

The golden age of medical AI research

AI routinely "beat" doctors on head-to-head evaluations

AI Outperforms Radiologists in Detecting Prostate Cancer on MRI Scans

By IPM staff June 14, 2024

Artificial Intelligence Outperforms Radiologists for Pancreatic Cancer Lymph Node Metastasis Prediction at CT

Linda C. Chu, Elliot K. Fishman

Author Affiliations

June 26, 2024

AI outperforms radiologists in detecting clinically significant prostate cancer

BREAST CANCER

AI outperforms radiologists in mammographic screening

David Killock

Nature Reviews Clinical Oncology **17**, 134 (2020) | Cite this article

9494 Accesses | **33** Citations | **292** Altmetric | Metrics

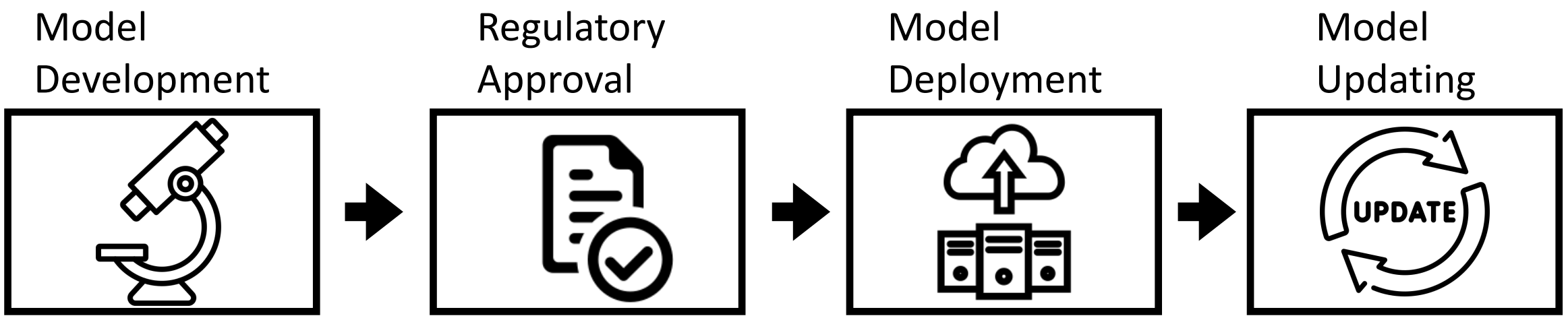
AI Outperforms Radiologists in Detecting Fractures in Patients At-Risk for Osteoporosis

March 30, 2020

By Whitney J. Palmer

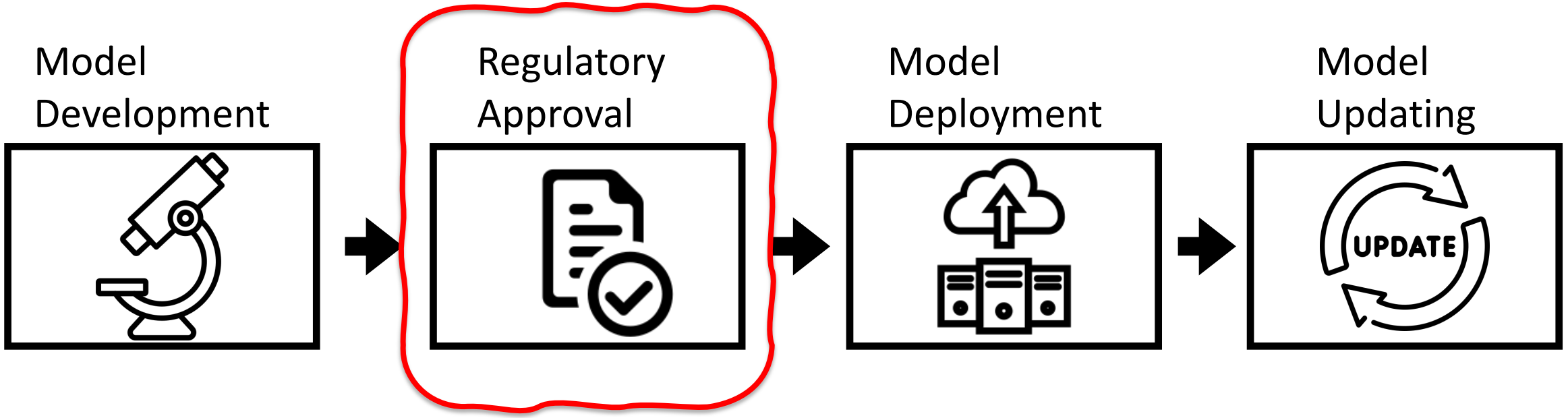
Medical AI Translation Pipeline

Changing healthcare with AI depends on more than just a good algorithm

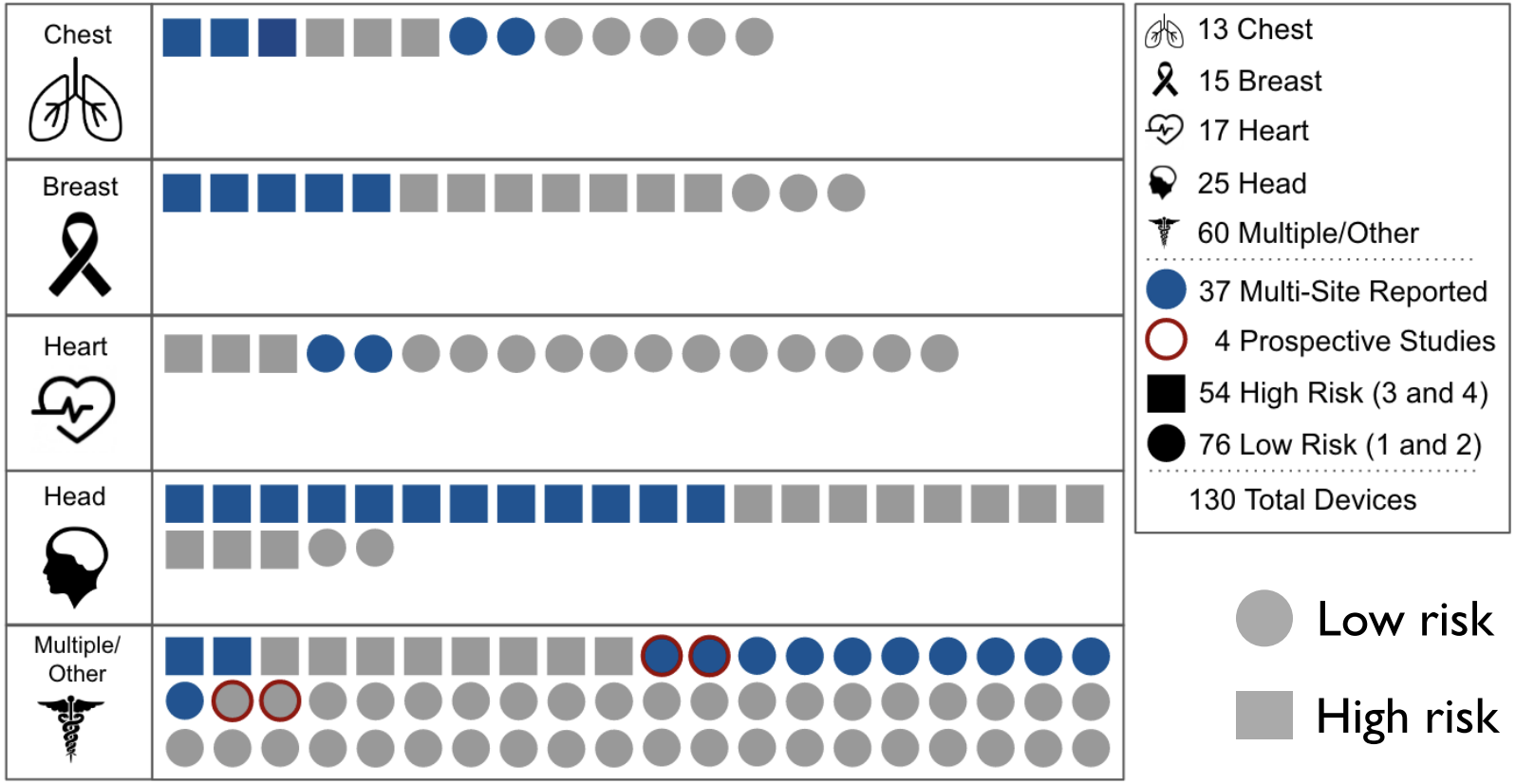


Medical AI Translation Pipeline

Changing healthcare with AI depends on more than just a good algorithm



Data used to evaluate 130 FDA-approved AI



93/130 did not report multi-site evaluation

Only 4 prospective studies

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals

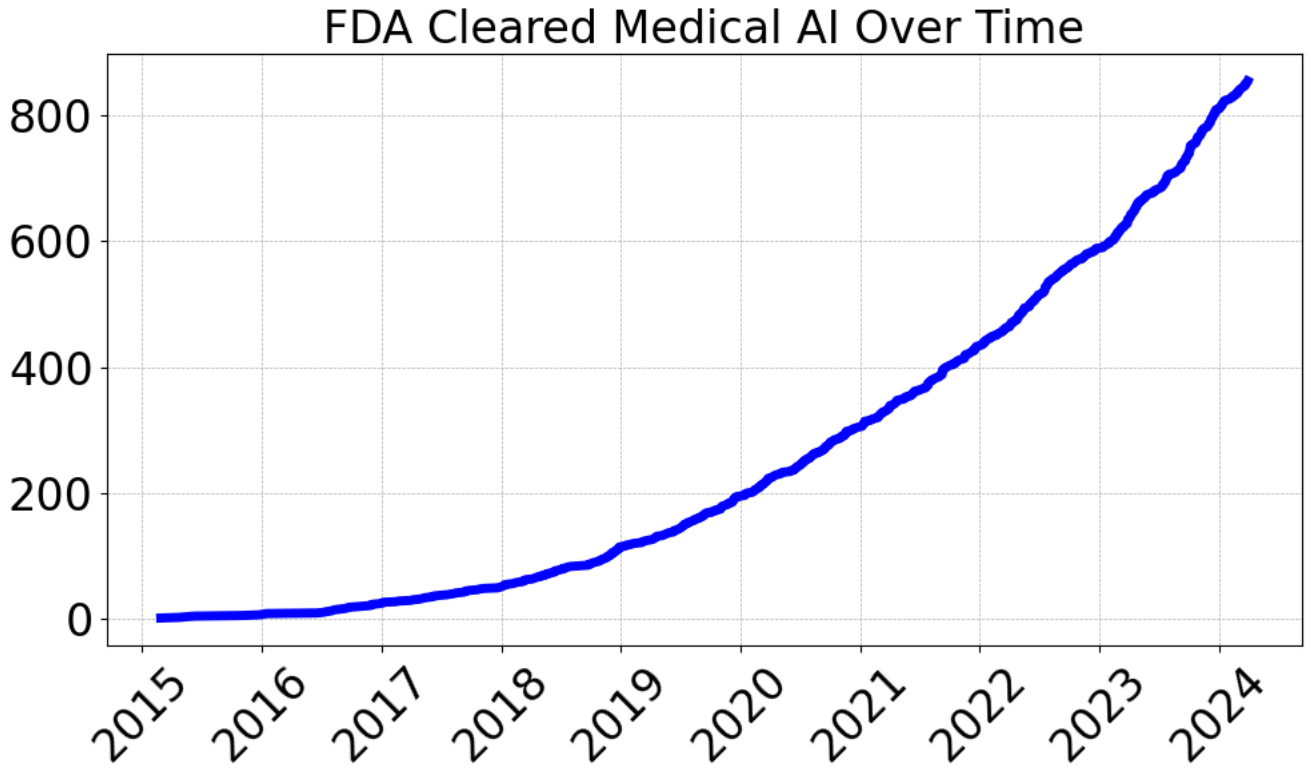
A comprehensive overview of medical AI devices approved by the US Food and Drug Administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho and James Zou

Wu et al. *Nature Medicine* 2021

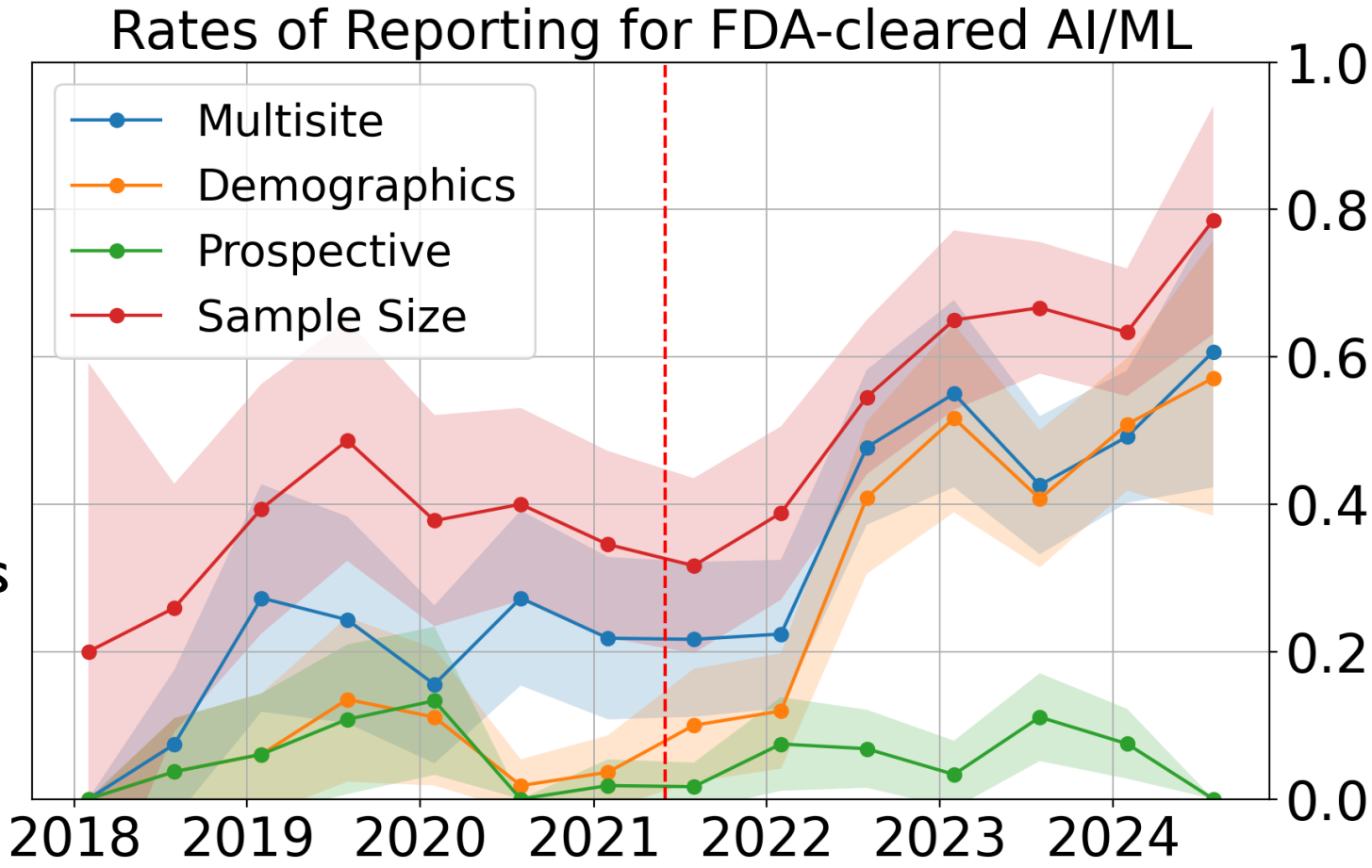
Rapid growth in the number of FDA cleared AI devices

- N=882 (and counting!)
- 151 new devices since Aug 2023



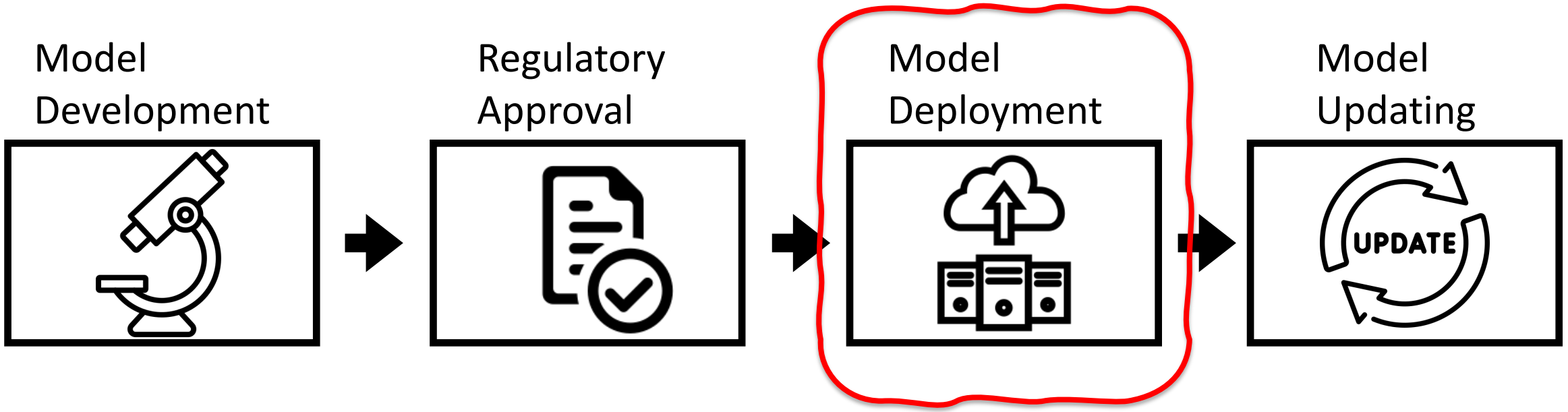
Some improvements in evaluation

- Significant increase in reporting after 2021
- Currently, more than half report multi-site evaluation and demographics
- Still very few prospective studies



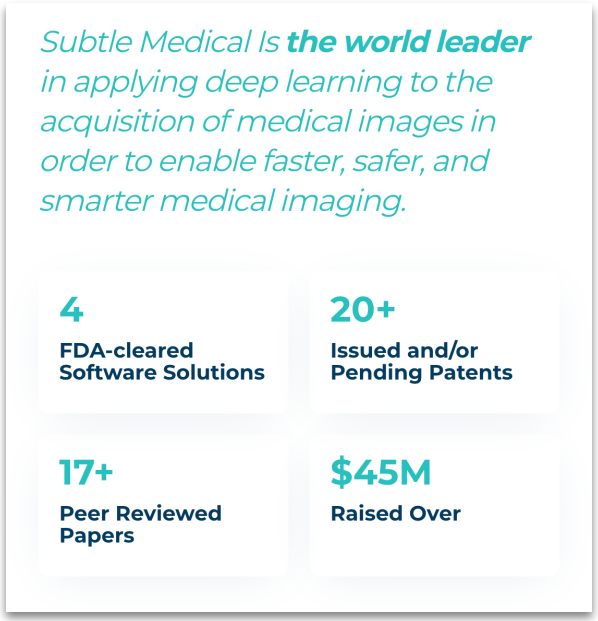
Medical AI Translation Pipeline

Changing healthcare with AI depends on more than just a good algorithm



AI Model Deployment

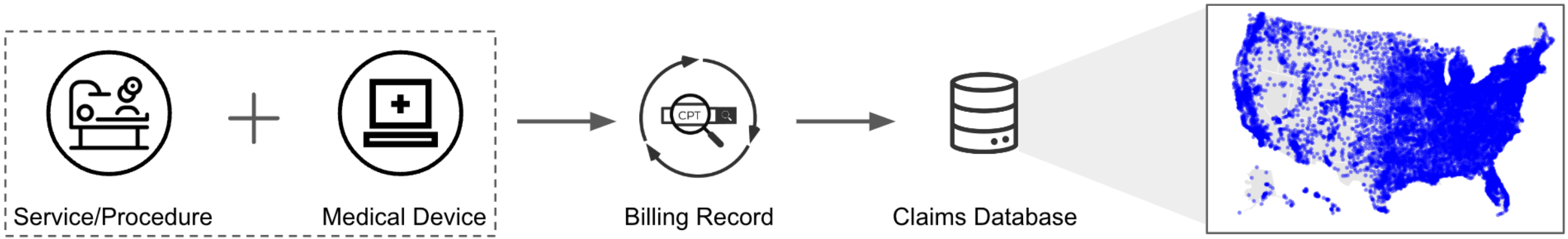
- Companies do not have to publicize how often their algorithms are used
- Real-world usage usually limited to press releases or case studies



Wu et al. *NEJM AI* 2024

Tracking AI deployment through insurance claims

Medical device usage is typically recorded through claims databases



Wu et al. *NEJM AI* 2024

Tracking AI deployment through insurance claims

Recently, there have been new CPT codes developed just for AI medical devices

Koios touts CPT codes for ultrasound AI decision-support software

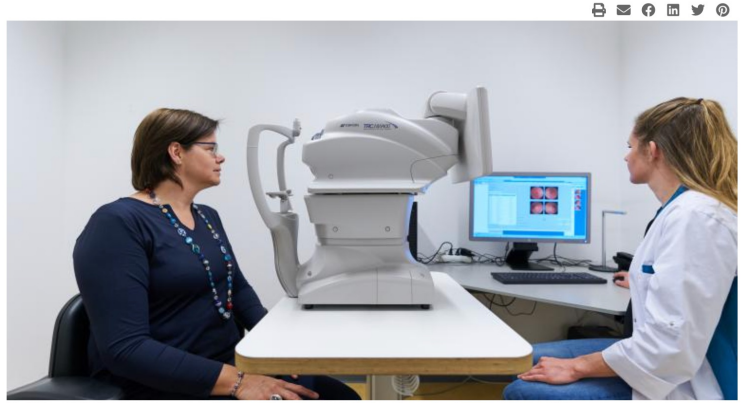
By AuntMinnie.com staff writers

November 4, 2022 -- New Current Procedural Terminology (CPT) codes issued by the American Medical Association (AMA) for ultrasound artificial intelligence (AI) software are generating reimbursements, according to Koios Medical.

CMS to reimburse providers for use of AI prostate cancer mapping tool

Hannah Murphy | *July 26, 2024* | *Health Imaging* | *Artificial Intelligence*

Autonomous AI Device Receives Historic CPT Code



IDx-DR, an FDA-cleared autonomous AI system that detects diabetic retinopathy. Image courtesy of IDx.

The AMA's Current Procedural Terminology (CPT) Editorial Panel has accepted a new Category 1 CPT code for retinal imaging with automated

Susan Shepard | Jul 31, 2019

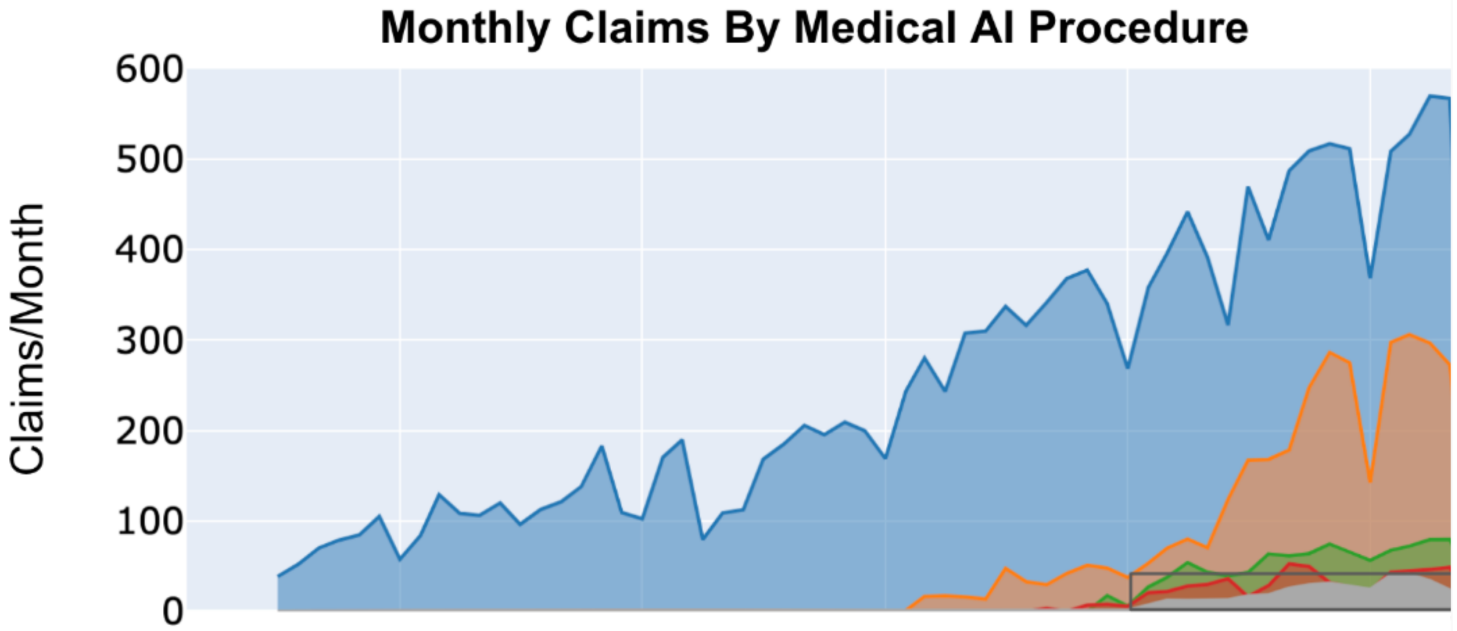
Wu et al. *NEJM AI* 2024

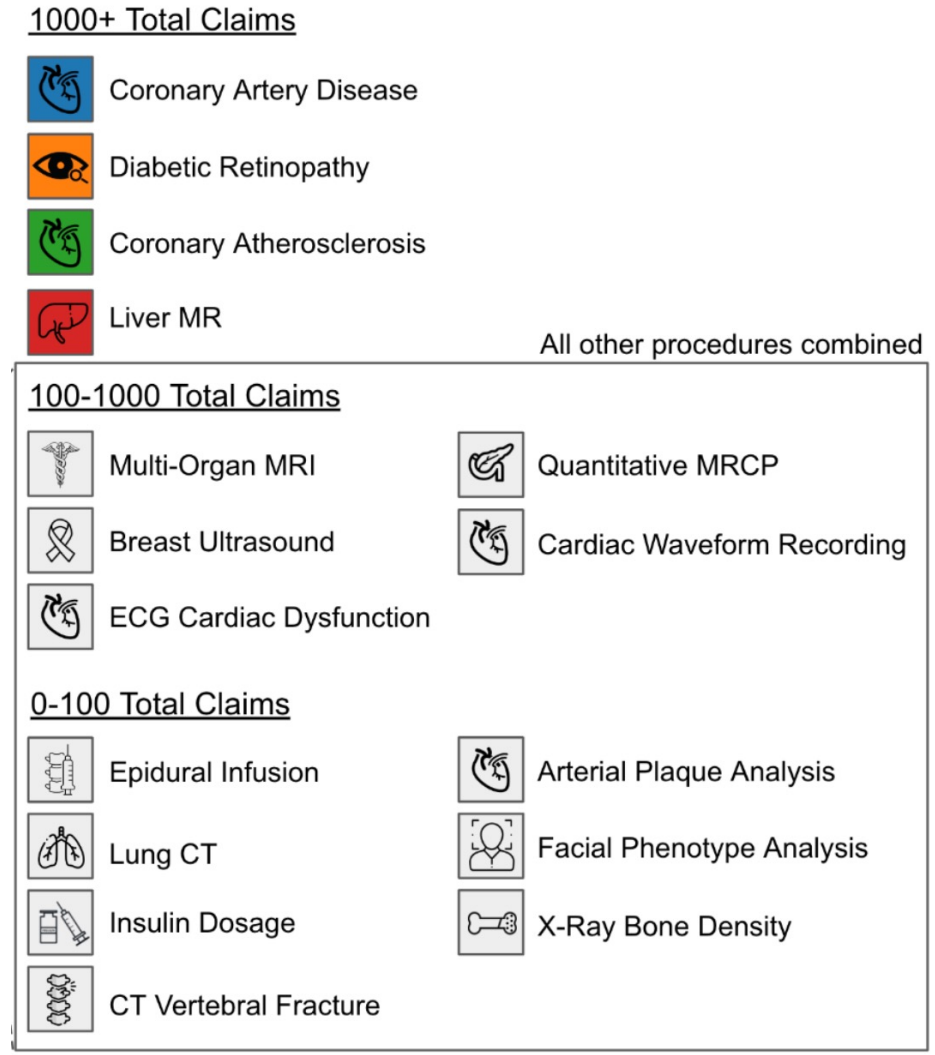
Only 4 AI devices with 1000+ claims

IQVIA PharMetrics® Plus

Comprehensive real world data on commercially insured patients in the US with a view into their diagnoses, treatments, and associated costs

- 210M unique patients
- 16B total claims from Jan 2015 -> June 2023 across 50 states
- Mostly private claims

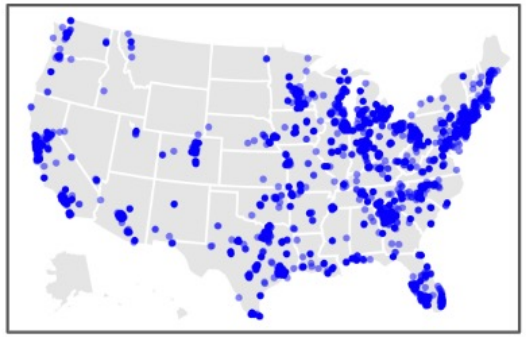




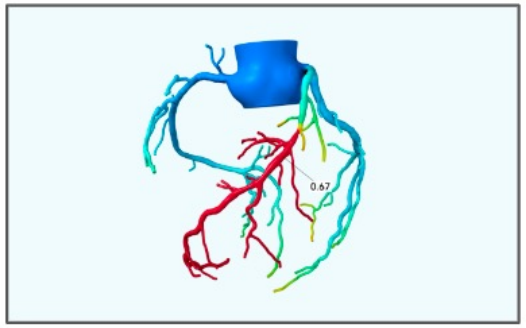
Wu et al. *NEJM AI* 2024

Examples products for top codes

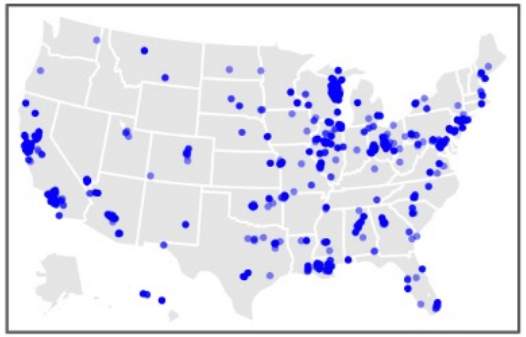
Coronary Artery Disease
0501T-0504T



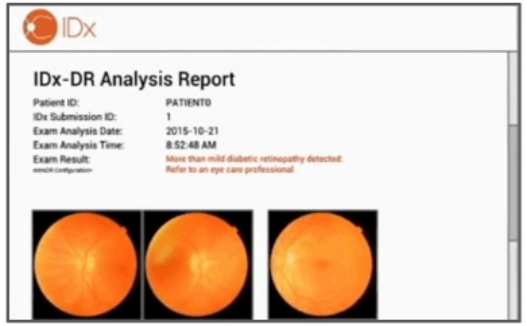
Example Product:
HeartFlow Analysis



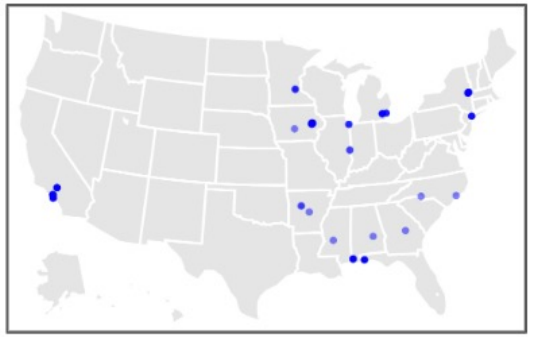
Diabetic Retinopathy
92229



Example Product:
LuminetricsCore

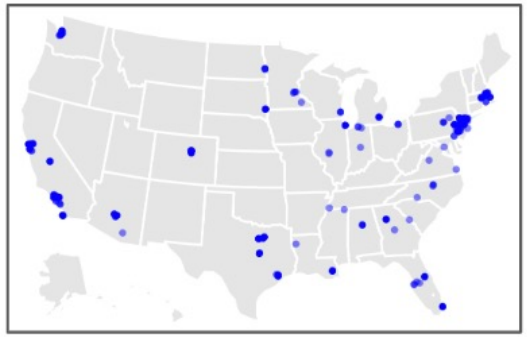


Coronary Atherosclerosis
0623T-0626T

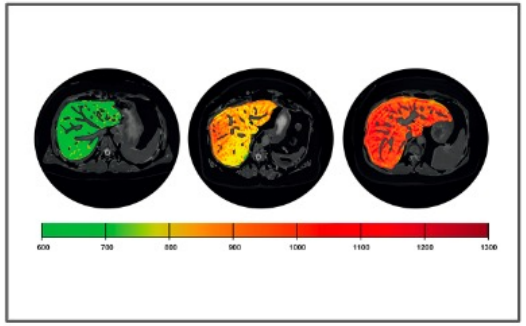


Example Product:
Cleerly

Liver MR
0648T, 0649T



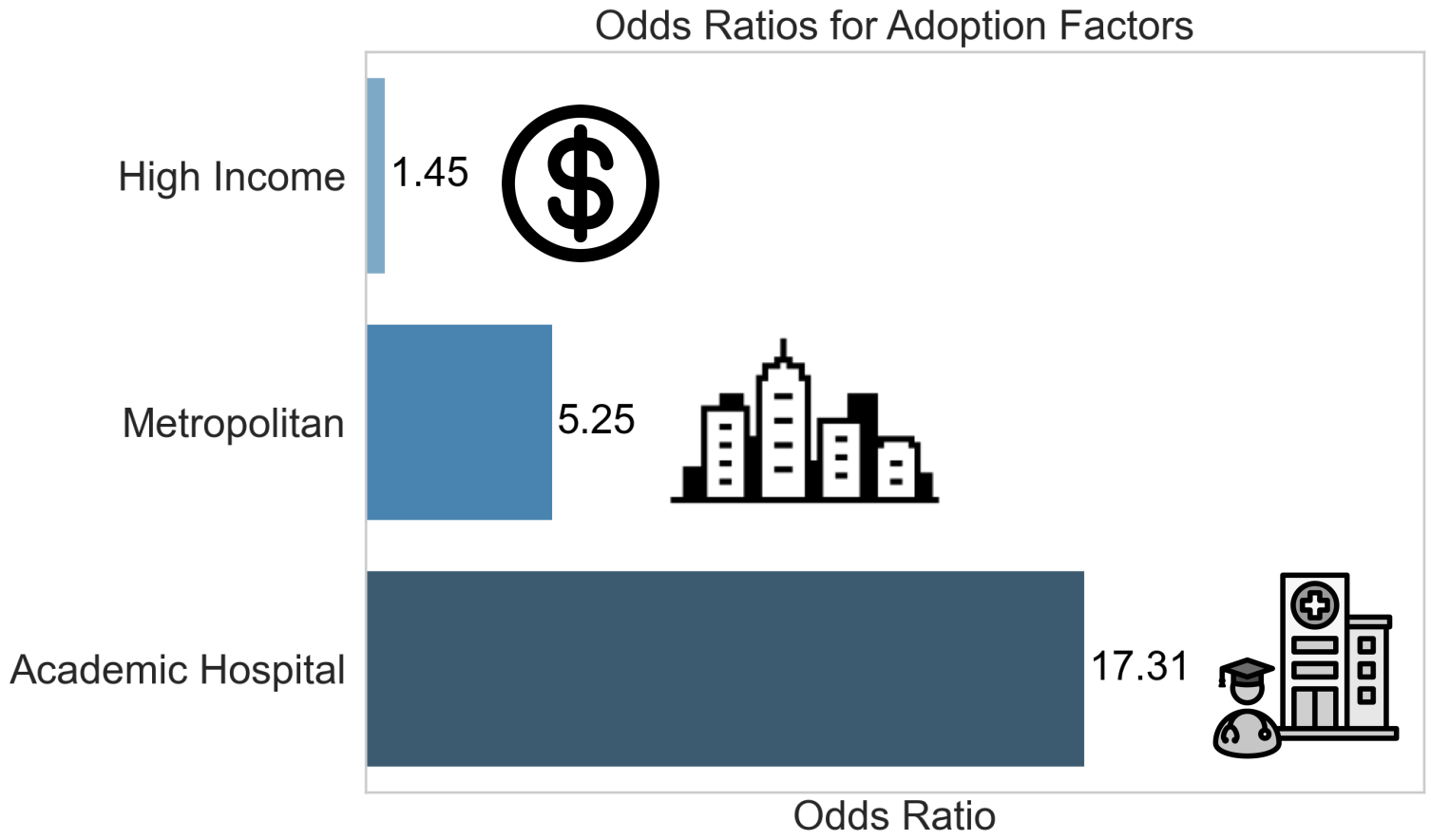
Example Product:
Perspectum LiverMultiScan



Wu et al. *NEJM AI* 2024

Factors driving AI adoption

- Academic hospitals are key drivers for AI adoption
- Rural zipcodes less likely to have AI
- Slight bias towards higher income zipcodes



Wu et al. *NEJM AI* 2024

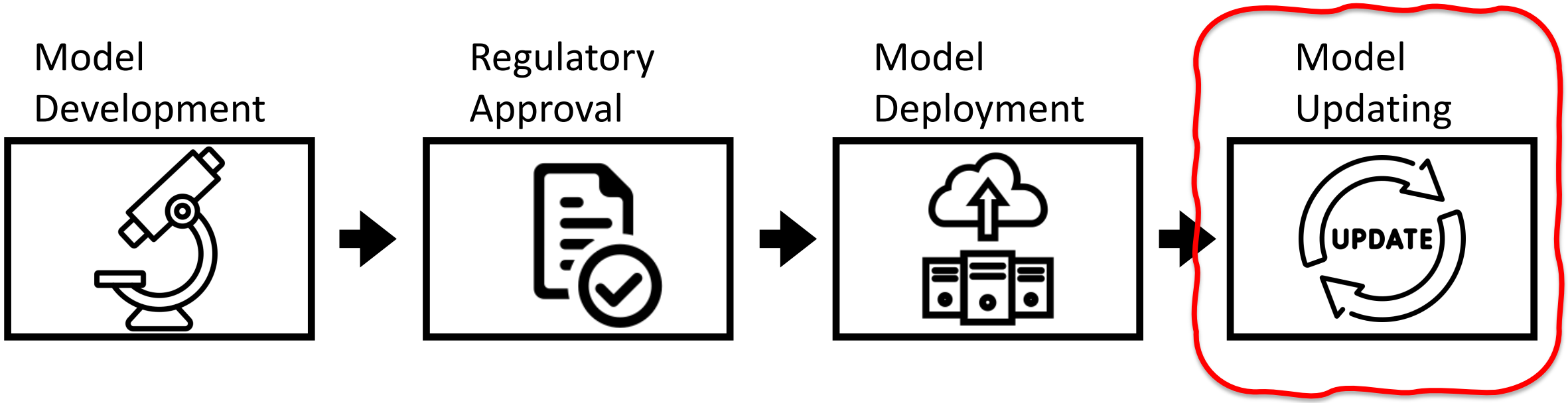
Key findings for medical AI deployment

- Medical AI adoption still nascent despite many FDA clearances
- Limited existing adoption follows typical tech trends:
  - High income, metropolitan, academic hospitals first
- Lack of clear payment mechanisms to drive adoption
  - CMS only pays for 1 CPT code currently

Wu et al. *NEJM AI* 2024

Medical AI Translation Pipeline

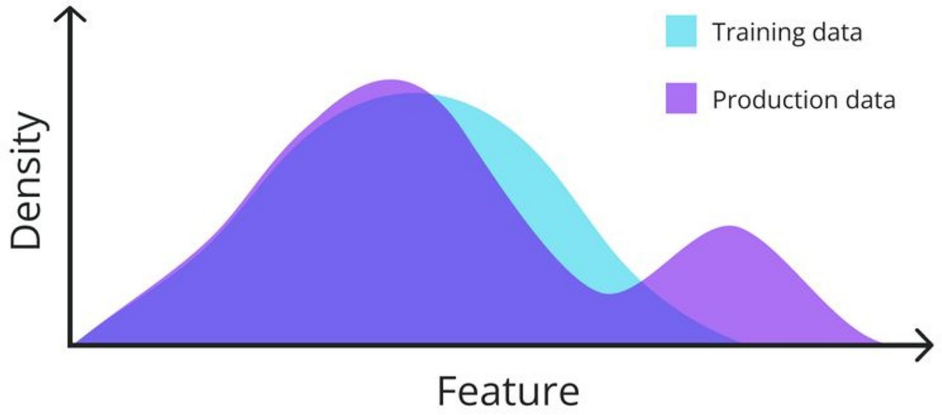
Changing healthcare with AI depends on more than just a good algorithm



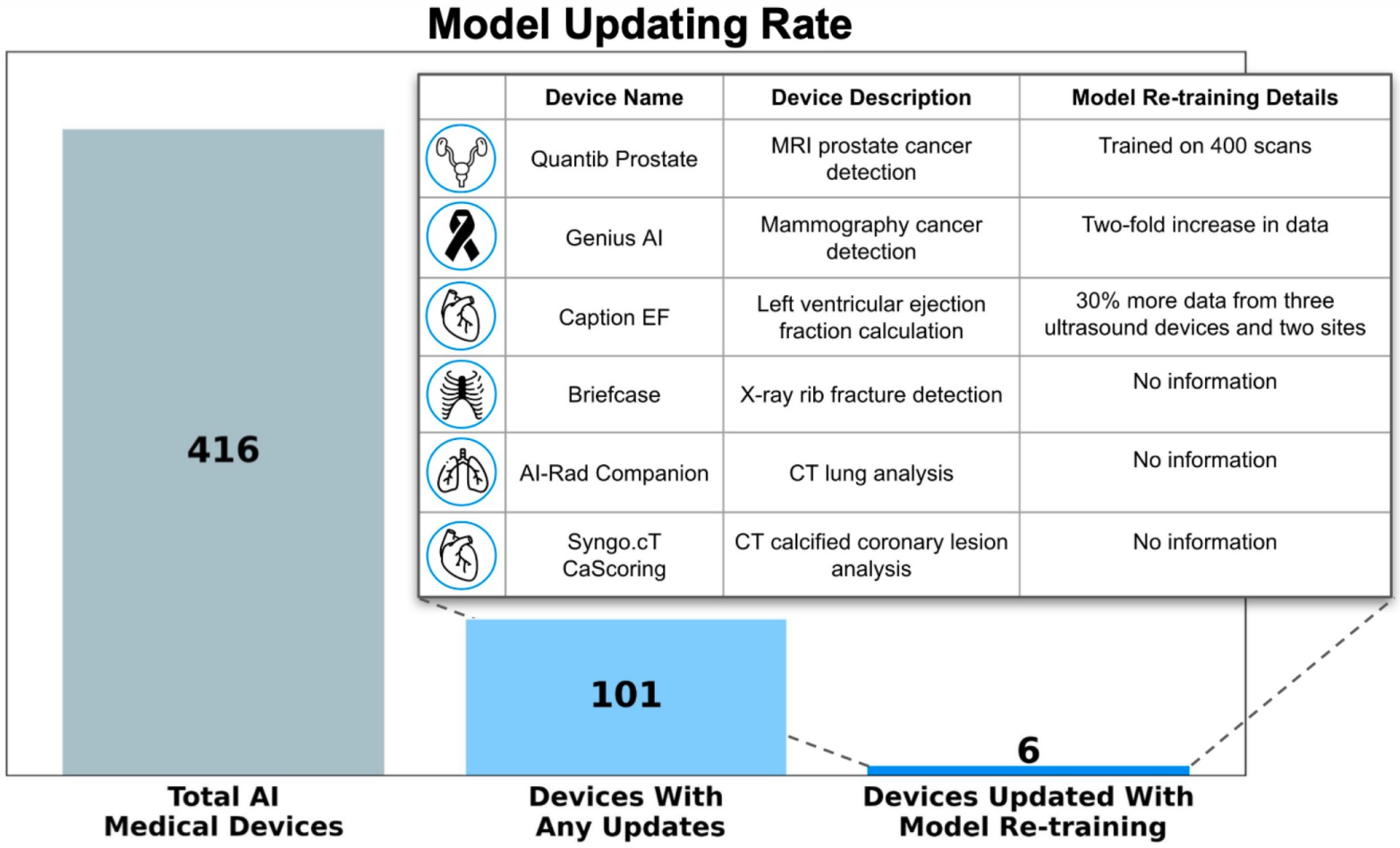
Medical AI Pipeline

A key promise of AI models is adaptability

- Addressing distribution shifts over time
- Updating model with architecture improvements
- Addressing uncovered biases



Medical AI algorithms rarely retrained



	Device Name	Device Description	Model Re-training Details
	Quantib Prostate	MRI prostate cancer detection	Trained on 400 scans
	Genius AI	Mammography cancer detection	Two-fold increase in data
	Caption EF	Left ventricular ejection fraction calculation	30% more data from three ultrasound devices and two sites
	Briefcase	X-ray rib fracture detection	No information
	AI-Rad Companion	CT lung analysis	No information
	Syngo.cT CaScoring	CT calcified coronary lesion analysis	No information

Wu et al. *CHIL* 2024

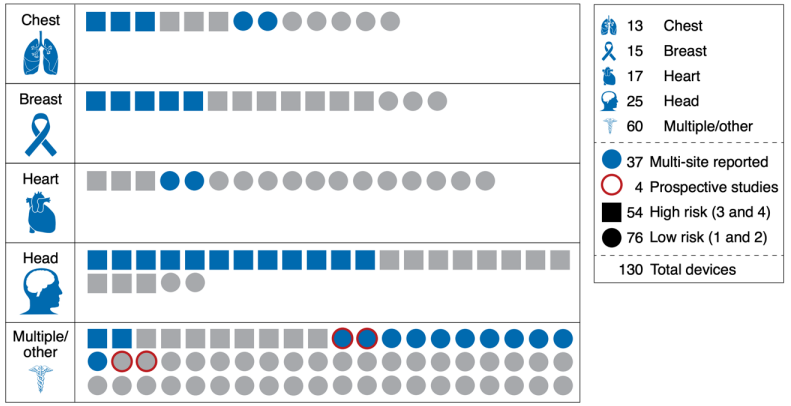
Medical AI updating

Implications

- Legacy FDA framework for device updates requires a new submission for any changes to the algorithm.
- Current FDA-cleared medical AI are commonly older versions
- FDA currently experimenting with Pre-determined Change Control Plan (PCCP)

Wu et al. *CHIL* 2024

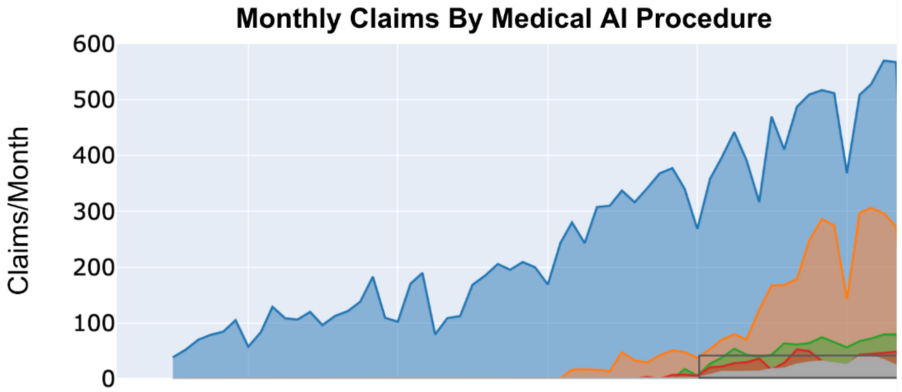
Key Takeaways

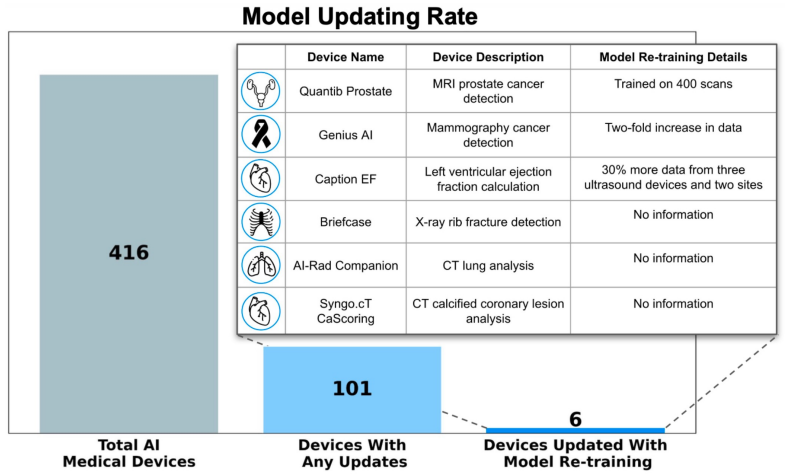


1. FDA's evaluation standards have evolved over time to adapt to changes in AI

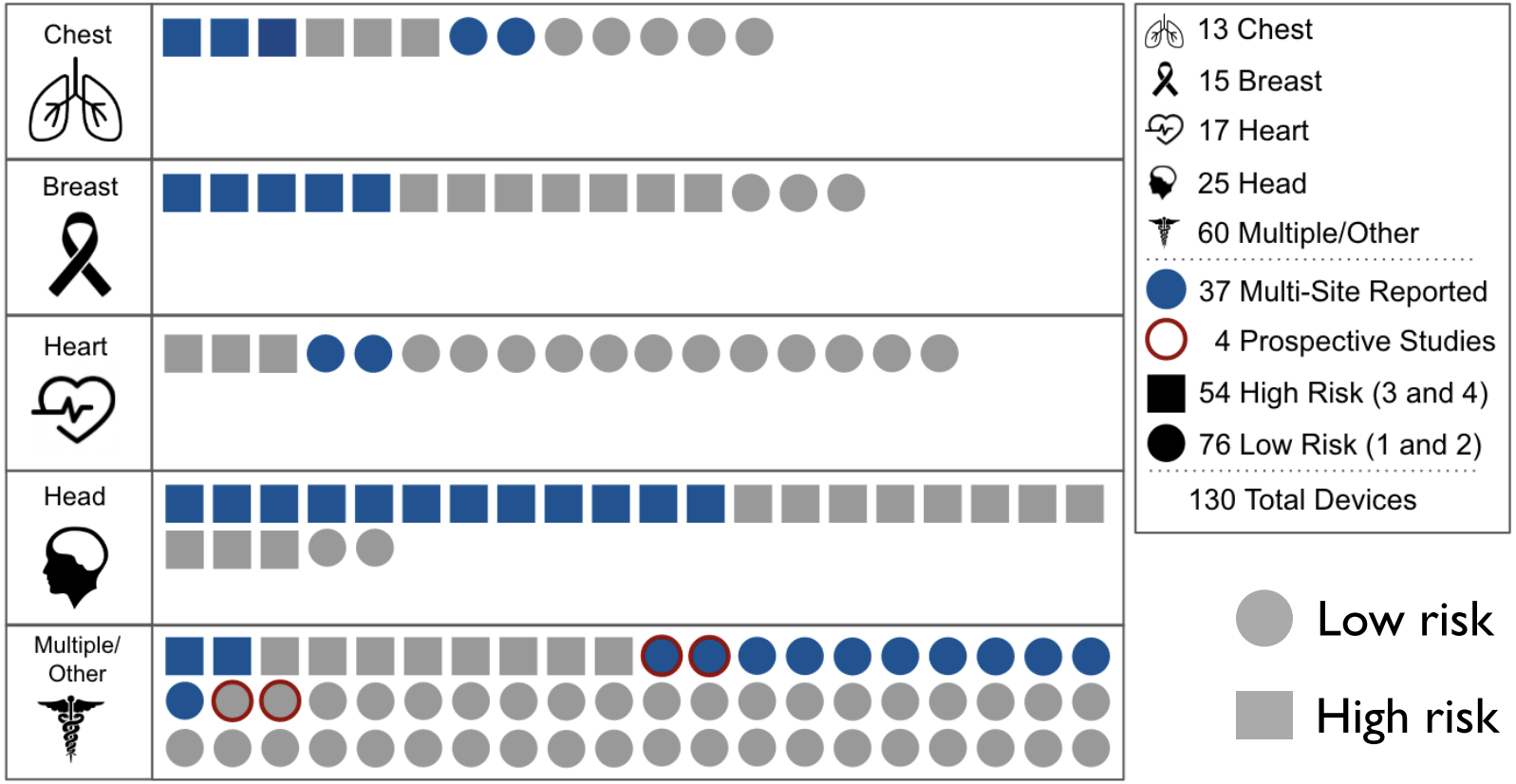
2. Medical AI needs to have a robust payment pathway for large-scale deployment

3. New framework needed to allow for updating in FDA-cleared medical devices





Data used to evaluate 130 FDA-approved AI



93/130 did not report multi-site evaluation

Only 4 prospective studies

How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals

A comprehensive overview of medical AI devices approved by the US Food and Drug Administration sheds new light on limitations of the evaluation process that can mask vulnerabilities of devices when they are deployed on patients.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho and James Zou

Wu et al. *Nature Medicine* 2021

Case study 1: EchoNet: from research to RCT

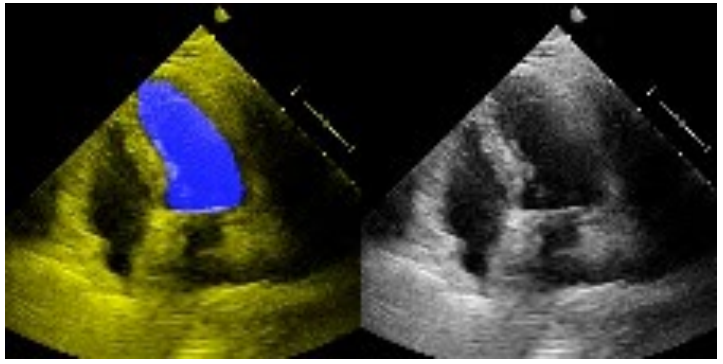
Article | Published: 25 March 2020

Video-based AI for beat-to-beat assessment of cardiac function

David Ouyang ✉, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley & James Y. Zou ✉

Nature **580**, 252–256(2020) | Cite this article

Ouyang et al. *Nature* 2020



David Ouyang

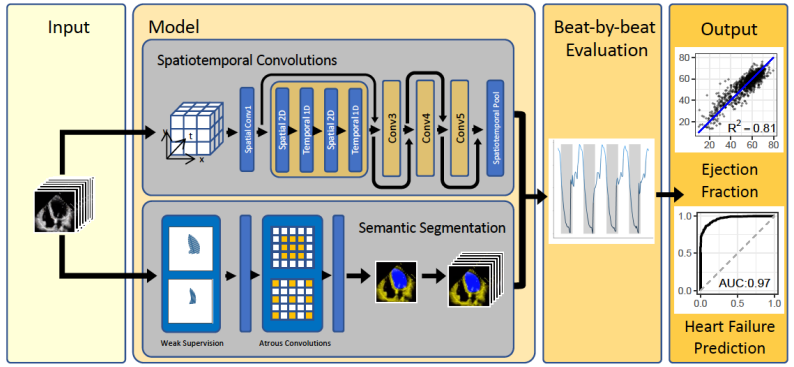
Article | Open Access | Published: 05 April 2023

Blinded, randomized trial of sonographer versus AI cardiac function assessment

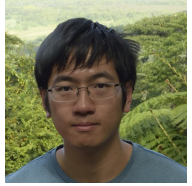
Bryan He, Alan C. Kwan, Jae Hyung Cho, Neal Yuan, Charles Pollick, Takahiro Shiota, Joseph Ebinger, Natalie A. Bello, Janet Wei, Kiranbir Josan, Grant Duffy, Melvin Jujjavarapu, Robert Siegel, Susan Cheng ✉, James Y. Zou ✉ & David Ouyang ✉

Nature **616**, 520–524 (2023) | Cite this article

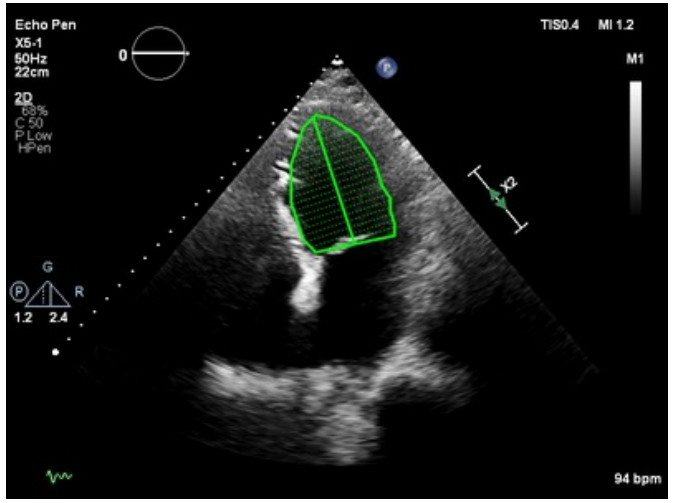
He et al. *Nature* 2023



Bryan He



Cardiac ultrasound videos

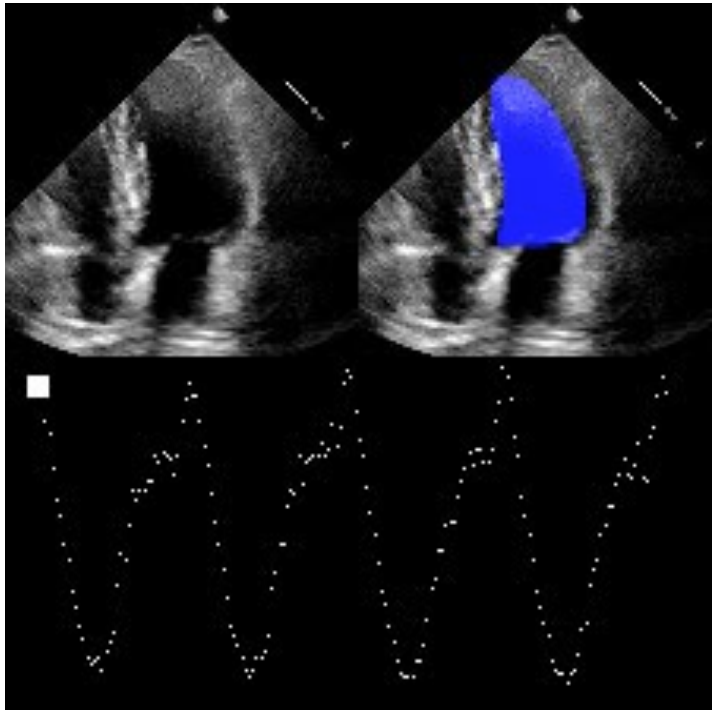


Echocardiogram is routinely used to assess heart function.
>10 million/year in the US.

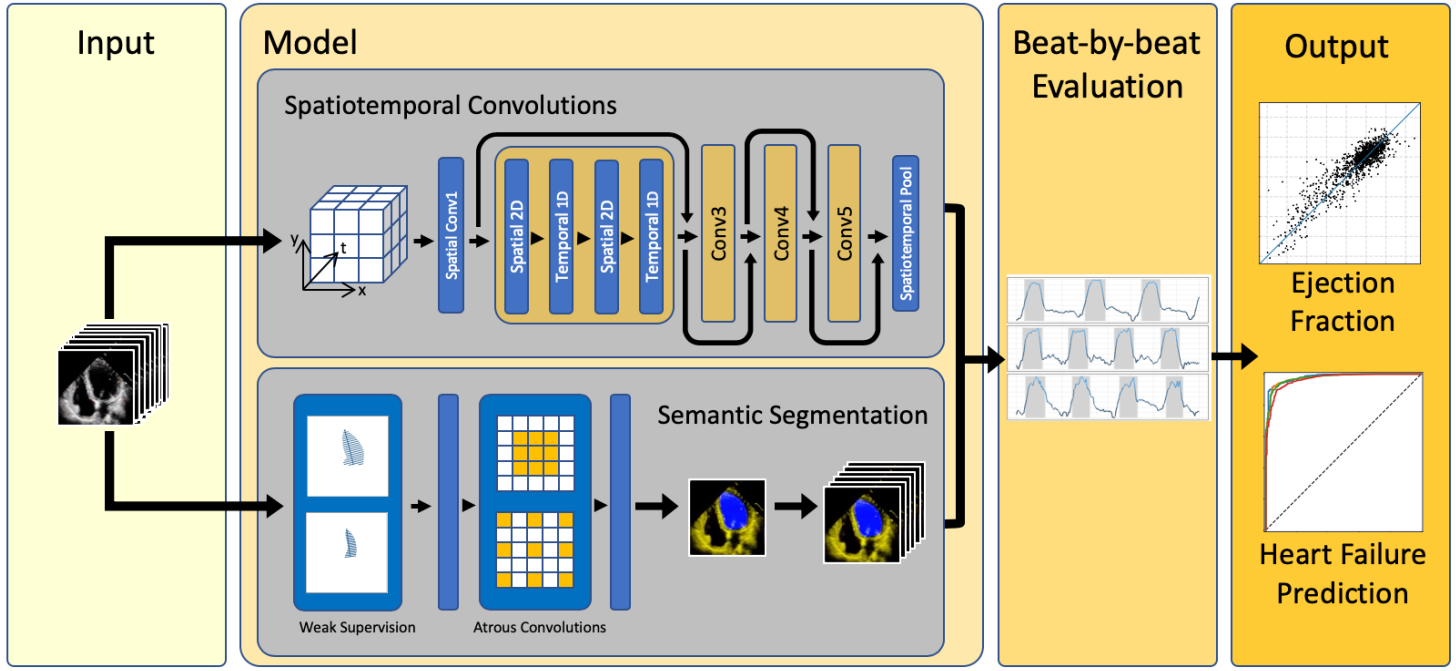
Standard clinical practices rely on manual segmentation: expensive, time-consuming and variable.

Computer vision can be faster and more reliable

EchoNet AI output



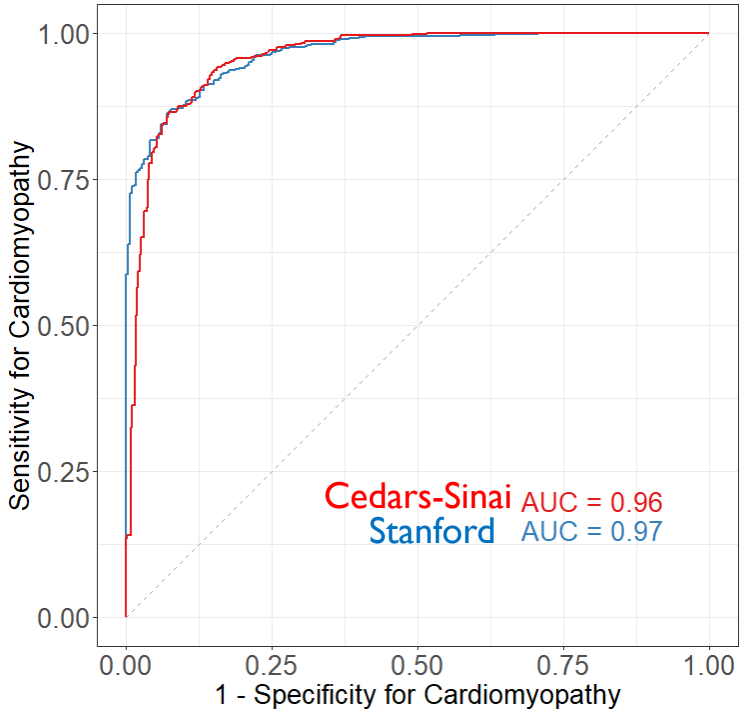
EchoNet architecture



Ouyang et al *Nature* 2020

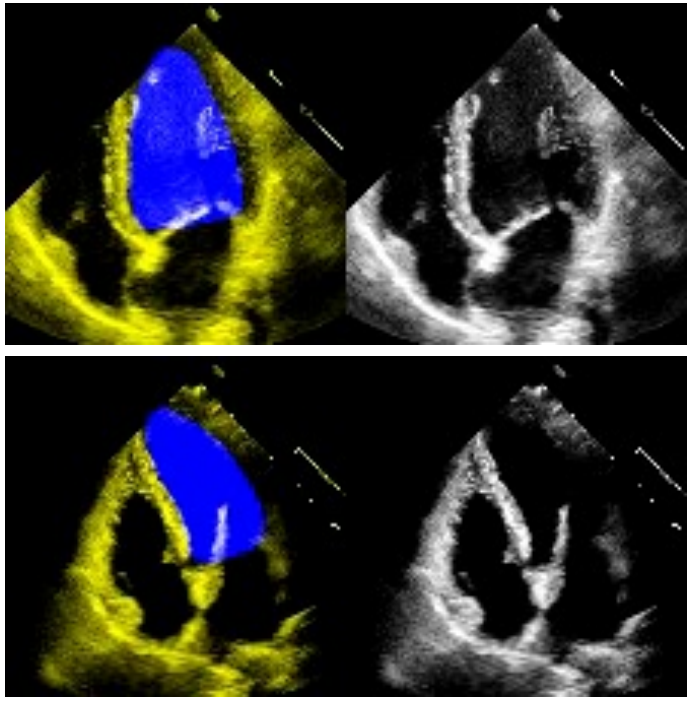
Retrospective evaluation: achieves expert performance

Predicting heart failure



Evaluation at two hospitals

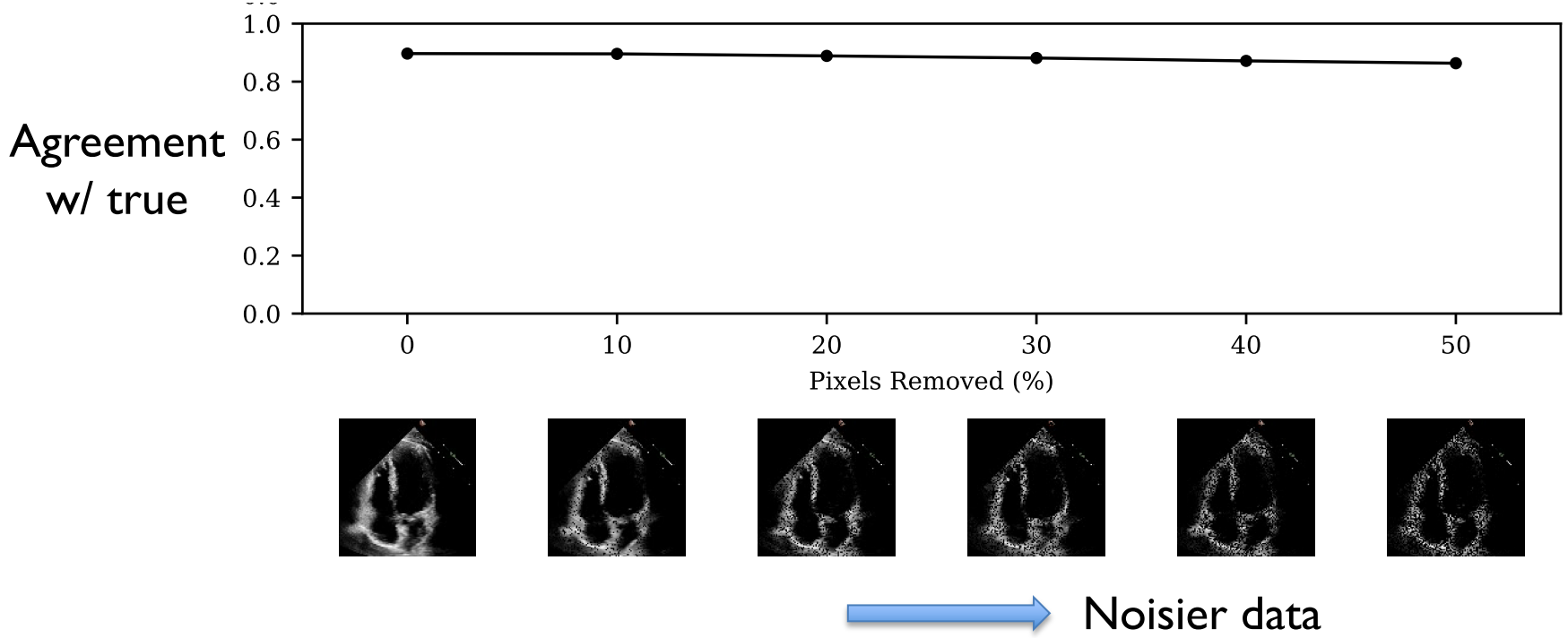
Examples



Ouyang et al *Nature* 2020

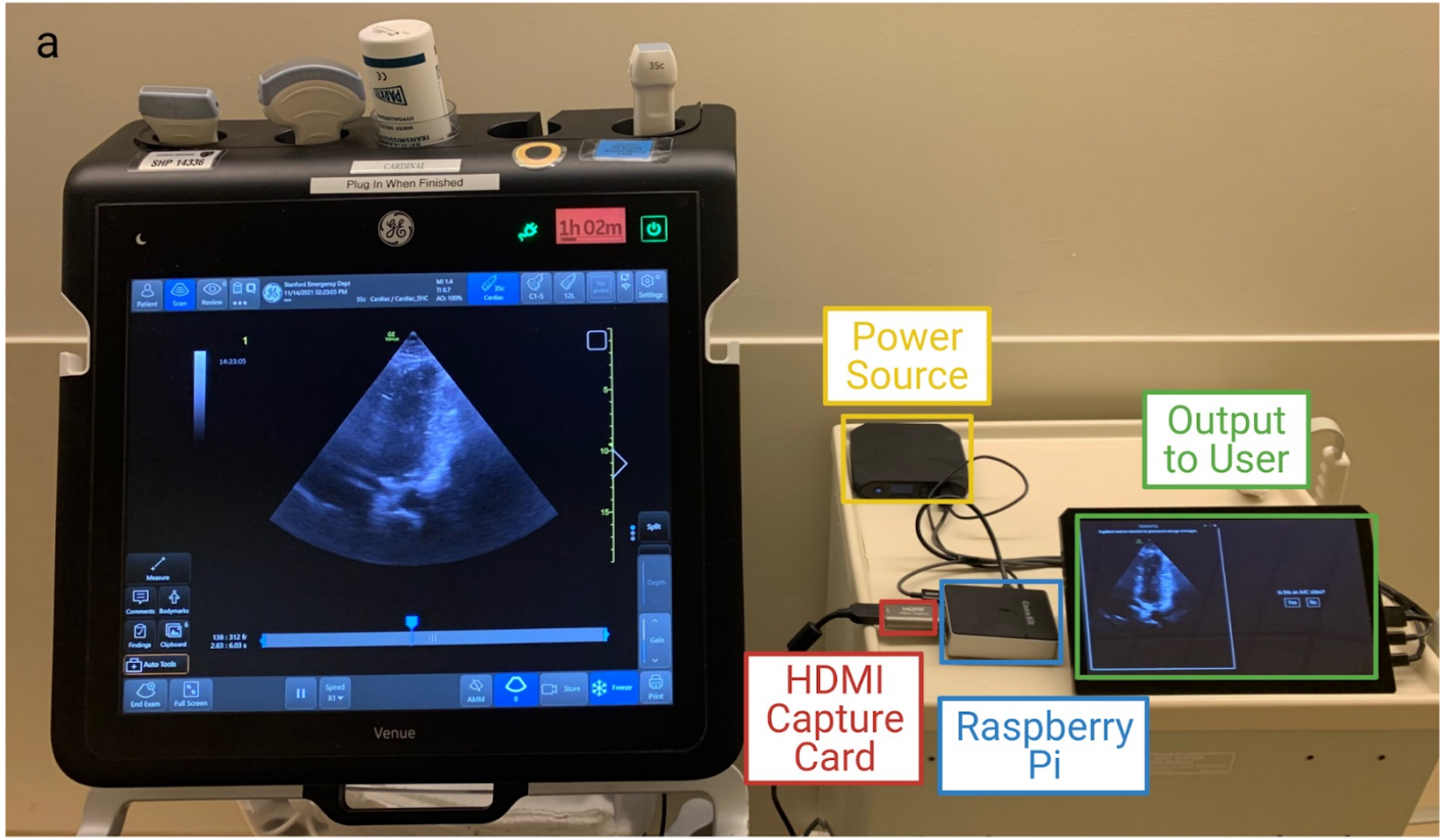
Algorithm robust to noisy videos

Maintains good performance when 50% of video is corrupted



Ouyang et al *Nature* 2020

Point-of-care application in emergency department



He et al. *J. Emerg. Med.* 2023

Randomized blinded clinical trial testing EchoNet

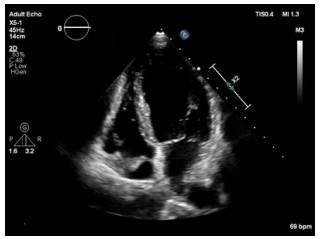
Article | Open Access | Published: 05 April 2023

Blinded, randomized trial of sonographer versus AI cardiac function assessment

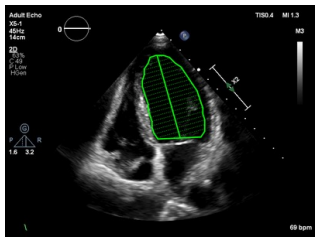
Bryan He, Alan C. Kwan, Jae Hyung Cho, Neal Yuan, Charles Pollick, Takahiro Shiota, Joseph Ebinger, Natalie A. Bello, Janet Wei, Kiranbir Josan, Grant Duffy, Melvin Jujjavarapu, Robert Siegel, Susan Cheng ✉, James Y. Zou ✉ & David Ouyang ✉

Nature **616**, 520–524 (2023) | Cite this article

Randomized blinded clinical trial testing EchoNet

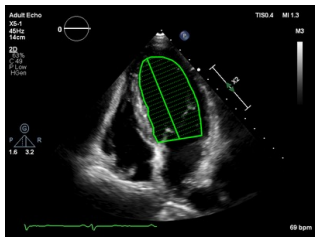


Sonographer Scan



AI Initial Assessment

Randomize 1:1



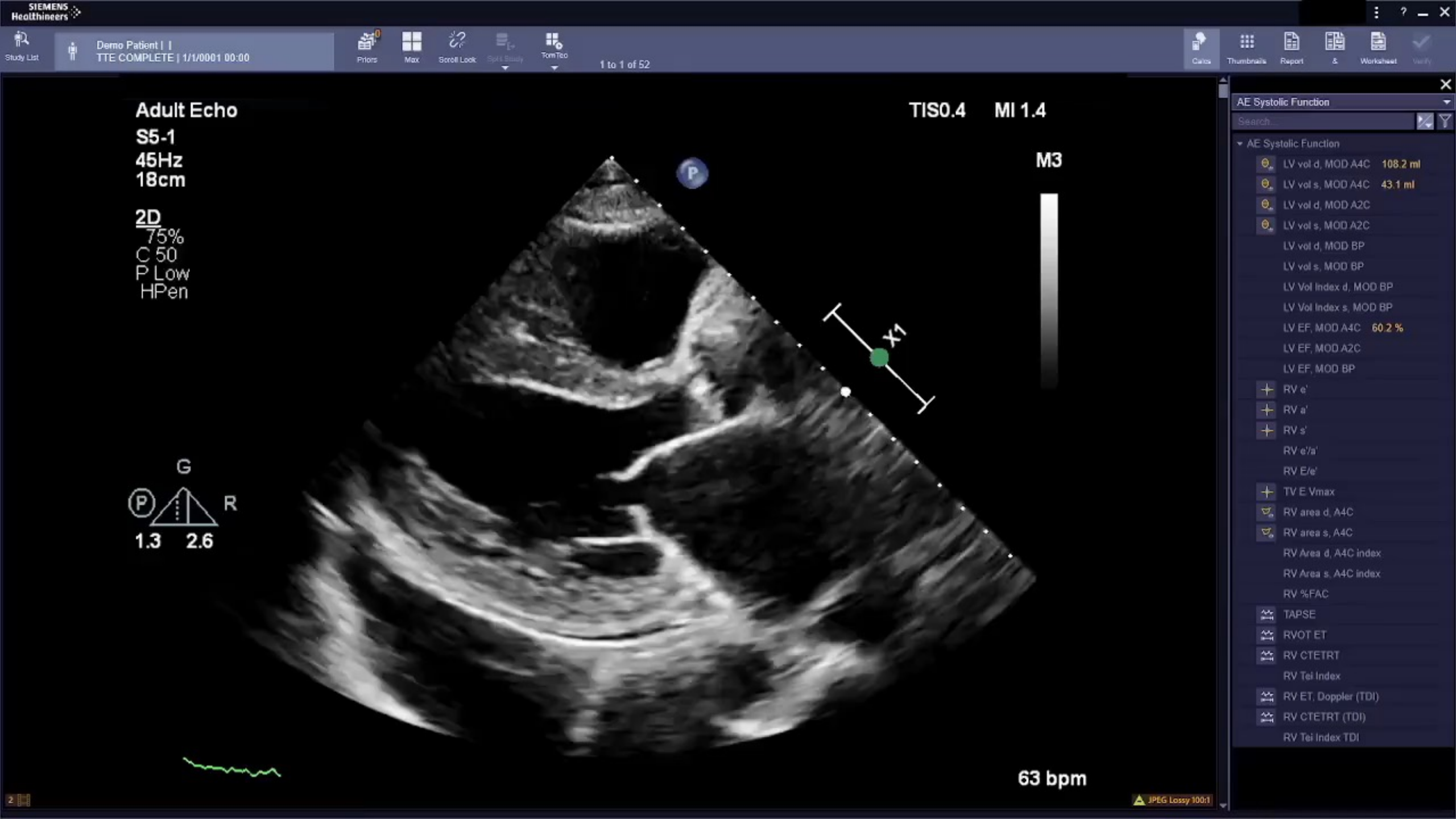
25 Human Sonographers Assessment

10 Blinded Cardiologist Report

Primary Outcome

Proportion of studies changed more than 5%

He et al. *Nature*. 2023


SIEMENS Healthineers
Study List
Demo Patient | |
TTE COMPLETE | 1/1/0001 00:00
Priors
Max
Scroll Lock
Split Study
TomTec
1 to 1 of 52
Calcs
Thumbnails
Report
&
Worksheet
Verify
Adult Echo
S5-1
45Hz
18cm
2D
75%
C 50
P Low
HPen
TIS0.4
MI 1.4
M3
X1
G
P
R
1.3
2.6
63 bpm
JPEG Lossy 100:1
AE Systolic Function
Search...
AE Systolic Function
LV vol d, MOD A4C 108.2 ml
LV vol s, MOD A4C 43.1 ml
LV vol d, MOD A2C
LV vol s, MOD A2C
LV vol d, MOD BP
LV vol s, MOD BP
LV Vol Index d, MOD BP
LV Vol Index s, MOD BP
LV EF, MOD A4C 60.2 %
LV EF, MOD A2C
LV EF, MOD BP
RV e'
RV a'
RV s'
RV e'/a'
RV E/e'
TV E Vmax
RV area d, A4C
RV area s, A4C
RV Area d, A4C index
RV Area s, A4C index
RV %FAC
TAPSE
RVOT ET
RV CTETRT
RV Tei Index
RV ET, Doppler (TDI)
RV CTETRT (TDI)
RV Tei Index TDI

Trial flow and patient characteristics

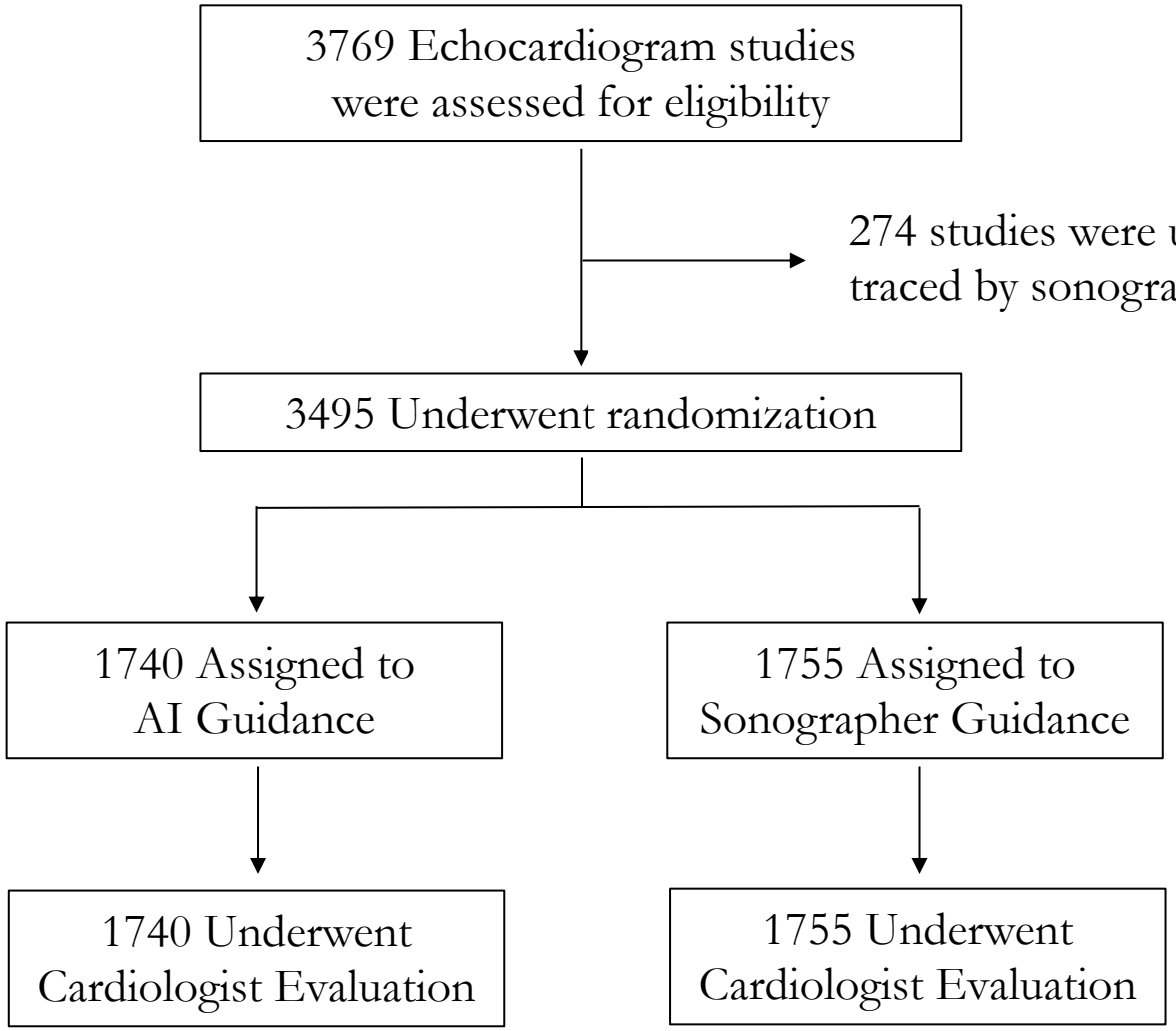
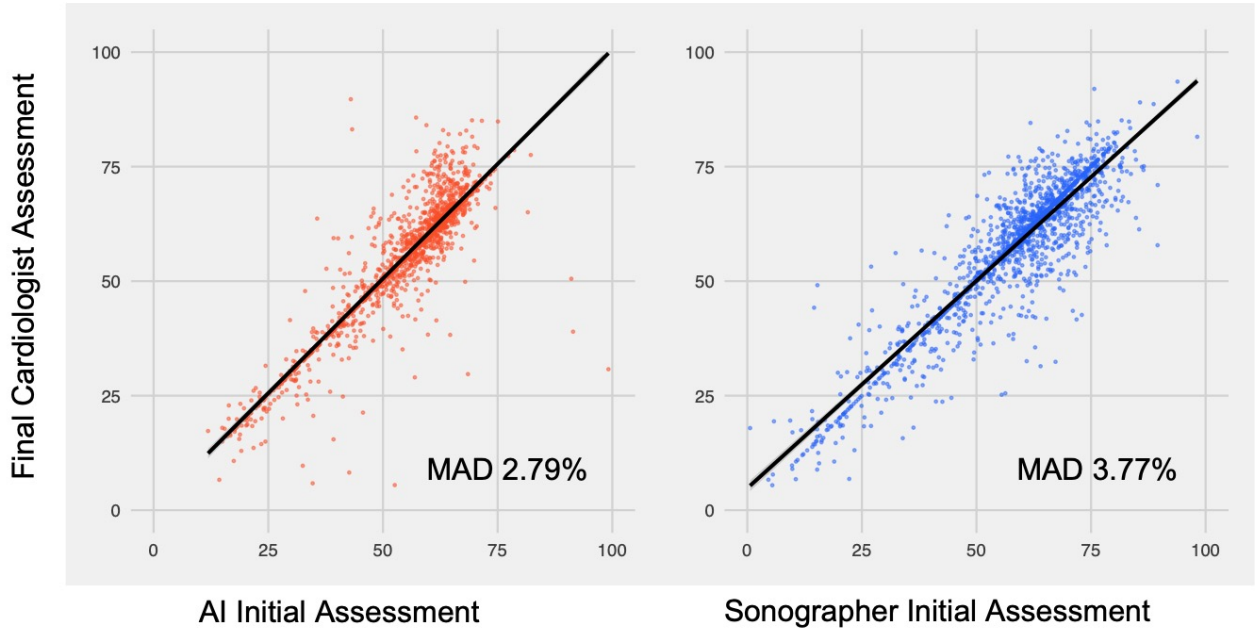


Table 1: Demographic and Imaging Study Characteristics.

Characteristic	Total (n = 3495)	AI (n = 1740)	Sonographer (n = 1755)
Age - yr	66.3±17.0	66.1±16.8	66.6±17.1
Sex - no. (%)			
Male	1983 (57%)	982 (56%)	1001 (57%)
Female	1512 (43%)	758 (44%)	754 (43%)
Race - no. (%)			
Non-Hispanic White	2041 (58%)	1032 (59%)	1009 (57%)
Black	479 (14%)	230 (13%)	249 (14%)
Hispanic	405 (12%)	203 (12%)	202 (12%)
Asian	273 (8%)	123 (7%)	150 (9%)
Other	237 (7%)	120 (7%)	117 (7%)
Unknown	38 (1%)	20 (1%)	18 (1%)
Pacific Islander	14 (0%)	8 (0%)	6 (0%)
American Indian	8 (0%)	4 (0%)	4 (0%)
Body Mass Index*	26.5±6.3	26.6±6.3	26.5±6.2
Comorbidities - no. (%)			
Hypertension	2019 (58%)	990 (57%)	1029 (59%)
Diabetes	884 (25%)	441 (25%)	443 (25%)
Coronary Artery Disease	1099 (31%)	547 (31%)	552 (31%)
Chronic Kidney Disease	882 (25%)	460 (26%)	422 (24%)
Atrial Fibrillation	867 (25%)	450 (26%)	417 (24%)
Prior Stroke	459 (13%)	225 (13%)	234 (13%)
Prior Clinical EF	58.1±14.3	58.1±14.2	58.0±14.4
Method of LVEF Evaluation - no. (%)			
Single Plane (A4C)	2249 (64%)	1107 (64%)	1142 (65%)
Biplane	1246 (36%)	633 (36%)	613 (35%)
Study Quality - no. (%)			
Poor	648 (19%)	314 (18%)	334 (19%)
Adequate	1725 (49%)	875 (50%)	850 (48%)
Good	236 (7%)	114 (7%)	122 (7%)
Not Specified	886 (25%)	437 (25%)	449 (26%)
Location - no. (%)			
Inpatient	2067 (59%)	1033 (59%)	1034 (59%)
Outpatient	1428 (41%)	707 (41%)	721 (41%)

A4C = Apical-4-Chamber, *BMI missing in 52 studies

AI improves ejection fraction assessment quality and efficiency



Outcome	AI (n = 1740)	Sonographer (n = 1755)	Mean Difference (95% CI)	P value
Primary Efficacy Outcome: Initial vs. Final Assessment				
Substantial Change	292 (16.8%)	478 (27.2%)	-10.5% (-13.2% to -7.7%)	< 0.001*
Mean Absolute Difference	2.79±5.53	3.77±5.22	-0.97 (-1.31 to -0.61)	< 0.001

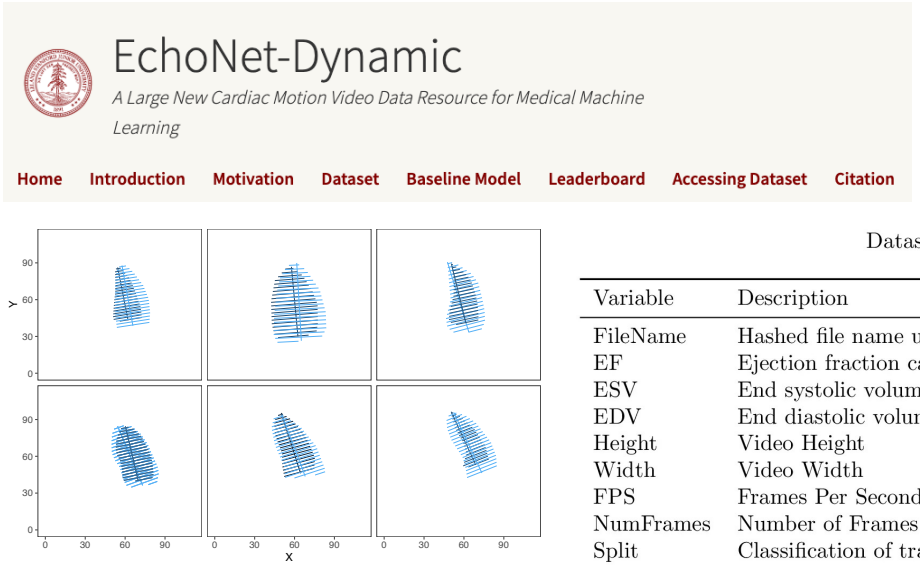
Successful blinding

Cardiologists could not distinguish between AI and sonographer initial assessments

Correct	1130	(32.3%)
Unsure	1520	(43.4%)
Incorrect	845	(24.2%)

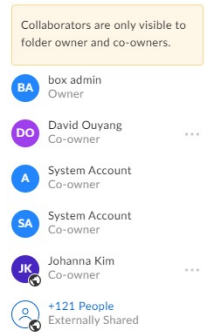
Outcome	AI (n = 1740)	Sonographer (n = 1755)	Mean Difference (95% CI)	P value
Primary Efficacy Outcome: Initial vs. Final Assessment				
Substantial Change	292 (16.8%)	478 (27.2%)	-10.5% (-13.2% to -7.7%)	< 0.001*
Mean Absolute Difference	2.79±5.53	3.77±5.22	-0.97 (-1.31 to -0.61)	< 0.001

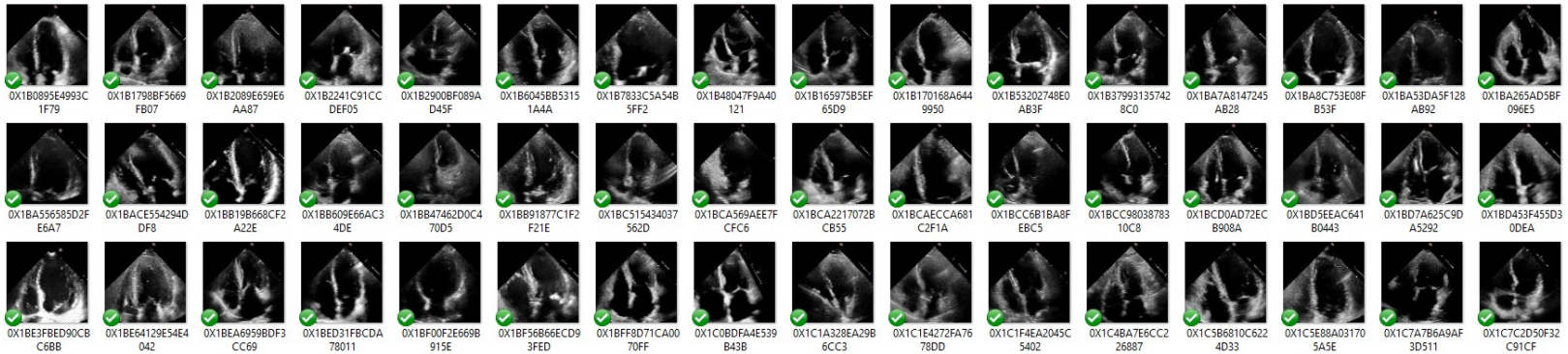
Open dataset and code



Dataset Label Variables

Variable	Description
FileName	Hashed file name used to link videos, labels, and annotations
EF	Ejection fraction calculated from ESV and EDV
ESV	End systolic volume calculated by method of discs
EDV	End diastolic volume calculated by method of discs
Height	Video Height
Width	Video Width
FPS	Frames Per Second
NumFrames	Number of Frames in whole video
Split	Classification of train/validation/test sets used for benchmarking



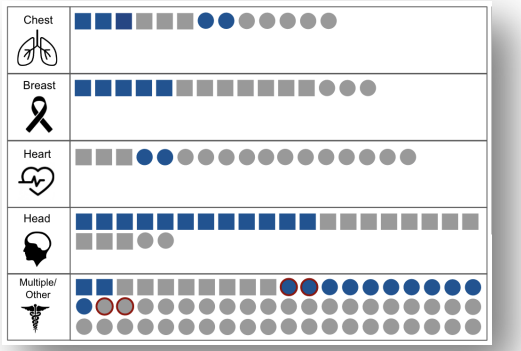


One of the largest public dataset of medical videos.

Ouyang et al *Nature* 2020; He et al *Nature* 2023

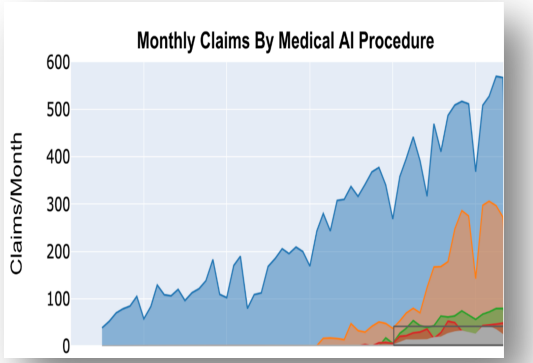
Medical AI evaluation and adoption

Medical AI regulatory evaluation



Wu et al *Nature Med* 2021

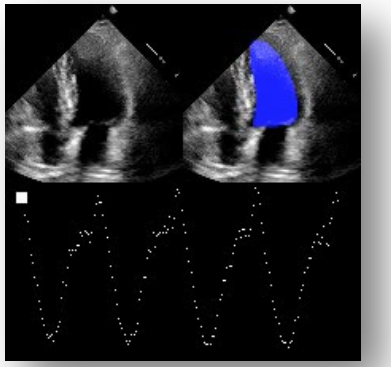
Medical AI adoption/updates



Wu et al *NEJM AI* 2024

Wu et al *CHIL* 2024

EchoNet AI FDA cleared



Ouyang et al *Nature* 2020

He et al *Nature* 2023

Thanks: Kevin Wu, Eric Wu, Daniel Ho, Michael Abramoff, Jimeng Sun, David Ouyang, Bryan He, IQVIA, CZ Biohub.